(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－134194

(43)公開日　平成９年(1997)５月20日

(51)Int.Cl.⁶	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ１０Ｌ　3/00	５７１		Ｇ１０Ｌ　3/00	５７１Ｇ	
Ｇ０６Ｔ　7/00			Ｈ０４Ｍ　11/00	３０２	
Ｈ０４Ｍ　11/00	３０２		Ｇ０６Ｆ　15/62	４６５Ｋ	

審査請求　未請求　請求項の数２　ＯＬ　（全６頁）

(21)出願番号　特願平8－209422

(22)出願日　平成８年(1996)８月８日

(31)優先権主張番号　０８／５１６，０９０

(32)優先日　1995年８月17日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号

(72)発明者　デイビッド　ジー　ストーク
アメリカ合衆国　カリフォルニア州　メンロー　パークスイート　115　サンド　ヒル　ロード　2882　リコー　コーポレーション　カリフォルニア　リサーチ　センタ内

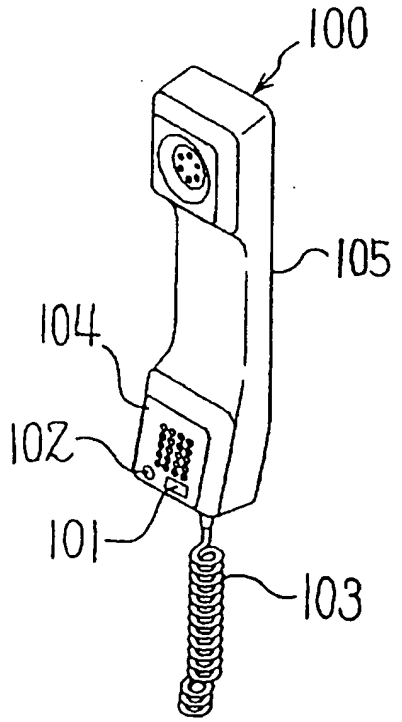
(74)代理人　弁理士：　柏木　明　（外１名）

(54)【発明の名称】　読話のための画像認識システム

(57)【要約】

【課題】　音声認識の認識精度を維持したまま、より多くの語彙についてより多くの人に適用できる音声認識を行うこと。

【解決手段】　送話器１０４、カメラ１０１、データ通信路１０３、及び認識処理論理回路を備える認識実行システムである。カメラ１０１は、送話器ハウジング１０５に取り付けられて読話のための少なくとも一つの顔の特徴に対応する画像情報を得る。得られるであろう顔の特徴は、舌の位置、歯の隙間、及び唇の丸くふくらむ出っ張り等である。このような画像情報は、データ通信路１０３を介して認識処理論理回路に送信され、画像情報の読話認識に供される。したがって、音響情報に基づく音声認識を補強するように画像情報に基づく読話認識を用いることで、認識精度を維持したまま、より多くの語彙についてより多くの人に適用できる音声認識を行うことができるようになる。



1

【特許請求の範囲】

【請求項1】　送話器ハウジングに収納された送話器と、

前記送話器ハウジングに取り付けられ、読話のための少なくとも一つの顔の特徴に対応する画像情報を得るカメラと、

このカメラに接続され、そのカメラから出力された画像情報を送信するデータ通信路と、

このデータ通信路に接続され、画像情報に基づく読話認識を実行する認識処理論理回路と、

を備えることを特徴とする読話のための画像認識システム。

【請求項2】　画像情報に基づく読話認識を実行する認識処理論理回路で処理されるデータを得るシステムであって、

送話器と、

この送話器に組み合わされ、読話のための少なくとも一つの顔の特徴に対応する画像情報を得るカメラと、

このカメラに接続され、そのカメラから出力された画像情報を前記認識処理論理回路に送信するデータ通信路と、

を備えることを特徴とする読話のための画像認識システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、音声認識の分野に係り、特に、映像音声認識のために顔の特徴の抽出を行う技術に関する。

【0002】

【従来の技術】自動的又は機械的な音声認識の最終目標は、話者のアクセント、性別、話す速度、発音の明瞭さの程度、聴覚上の騒雑音として存在している全てのものの等の障害を超えて、人間が理解するように音声を認識することである。このような音声認識の代表的なアプローチとしては、変化率が音素要素を表現するノード（ステート）間のリンク中で符号化される隠れマルコフモデル、ニューラルネットワークによる方法、複合的な特別の目的を持った音韻的、辞書的及び文法的な基礎を持つサブシステムが結合して協働し、音声認識のスコアを最大限にする「ブラックボード」の方法等がある。しかし、これらの各アプローチをシステム化した音声認識に関する近年のシステムでは、コンピュータによる発話文書間変換や自動翻訳等の応用分野で求められる数多くの要求を満たすのに必要な正確さ及び強健さが共に十分ではない。

【0003】従来、文法及び構文上のデータのような高いレベルの言語情報を認識処理に含めることに多くの研究が集中している。これに対し、人工的な音声認識システムに合理的に組み入れることができる情報予測や制約といったものも、音声認識の正確さを高める方向に働く。例えば、人間は、聴覚で捕えた音声の理解を増すために、音響信号以外の情報、例えば視覚的な情報を利用することがあり、このような予測情報も音声認識の認識率を高める。これは、聴覚が害された人が視覚的な情報を活用して音声を正しく理解することがしばしば見受けられることからも明かであろう。このような視覚情報を利用した音声認識としては、

（a）Doxld,B. and Campbell,R.(eds.),「目によるヒアリング：読唇の真理(Hearing by Eye : The Psychology of Lipreading)」,N.J.,Lawrence Erlbaum Press(1987)

（b）DeFilippo,C.L. and Sims,D.G.(eds.),「読話に関する新たな意見(New Reflections on Speechreading)」, special issue of The Volta Review 90(5),(1988)

を参照されたい。

【0004】話をしている人の視覚的情報に基づき音声を認識する読唇、読話では、音節及び音素について直接的な情報が得られる。話す速度、話者の性別、話者の同一性、背景の雑音から音声を分離するための微妙な情報も同様に得られる。このため、音声情報が多くの雑音によって崩れてしまう周知の「カクテルパーティー効果」が生じていても、話者の顔を見ることができる場合にはその人の話の内容がより明瞭になる。これは、音声認識に際して人間が視覚情報を用いることの強力な証拠となろう。

【0005】近年、以下のような多くの読話システムが記述されている。

（a）Petajan, E.D., 他,「音声認識を高めるための進歩した自動読唇システム(An Improved Automatic Lipreading System to Enhance Speech Recognition)」,ACM SIGCHI-88, 19-25(1988);

（b）Pentland,A., 他,「読唇：発声単語の自動視覚認識(Lip Reading : Automatic Visual Recognition of Spoken Words)」,Proc. Image Understanding andMachine Vision, Optical Society of America, June 12-14(1984)

（c）Yuhas,B.P., 他,「ニューラルネットワークを使用する音響及び視覚発話信号の統合(Integration of Acoustic and Visual Speech Signals Using Neural Networks)」, Nov.1989, IEEE Communications Magazine(1989)

Petajan,他は、視覚認識のためのラベル付けされた発声及び標準的な距離の分類を予め格納した辞書に合わせた言葉の生成の間、話者の顔の映像（画素）を使用することを記述している。Pentland, 他は、口を映すビデオ画素映像から上唇、下唇、及び口の二箇所の角部の速度を見積もるための光学的な流れの技術を記述している。そして、彼らは、3又は4桁の句について、構成要素分析及び最小距離分類の原理を使用する。Yuhas, 他 は、異

2

なるレベルの音響的雑音が存在している中で最良の認識を得るための視覚的及び音響的原因についての関連する重みを調整するためのフリーパラメータを伴うコントローラと共に、母音認識のための口の輪郭の静的な映像を使用するニューラルネットワークのトレーニングについて論じている。

【0006】他の典型的な読話システムとして、次のものがある。

（a）米国特許4，975，960，1990年12月4日発行，「電子的な顔の追跡及び検出システム及び自動化された音声認識のための方法及び装置(Electronic Facial Tracking and Detection System and Method and Apparatus forAutomated Speech Recognition)」(Pentajen)

（b）D.Stork, V.Prasad, G.Wolff,「読話の人間及び機械学習(Human and Machine Learning of Speechreading)」, the Computational Learning and NeuralLearning Workshop, Provincetown, Mass.に提出, September, 1993

（c）Stork, Wolff, and Levine,「向上した音声認識のためのニューラルネットワーク読唇システム(Neural Network Lipreading System for Improved Speech Recognition)」, IJCNN Int'l Joint Conf. on Neural Networks, IEEE(NewYork, NY), 1992, pgs.289-95 vol.2

（d）P.Silsbee & A.Borik,「自動読唇(Automatic Lipreading)」,30th Int'lBiomedical Sciences Instrumentation Symposium, vol.29, pgs.415-422(1993)

【0007】

【発明が解決しようとする課題】自動音声認識の適用技術については、例えば電話システムが大きな市場シェアを占めている。その一例として、株や商品の販売会社における電話注文による自動取引システムがある。これは、電話で話す顧客の音声を認識し、その指示に基づいて株や商品を自動売買するような電話システムである。このようなシステムでは、周囲の雑音に対して、個々の人が話した情報が正確かつ高い信頼性で録音、再生されることが最も重要であり、現在ではある程度の成功を収めている。

【0008】ところが、電話注文による自動取引システム等の電話システムが音声認識の認識率について成功を収めているとしても、そのようなシステムではほんの僅かばかりの語彙を利用できるに過ぎないし、利用できる話者も限定されてしまう、という問題をがある。

【0009】

【課題を解決するための手段】請求項1記載の発明は、読話のための画像認識システムであり、送話器ハウジングに収納された送話器と、送話器ハウジングに取り付けられて読話のための少なくとも一つの顔の特徴に対応する画像情報を得るカメラと、このカメラに接続されてカメラから出力された画像情報を送信するデータ通信路と、このデータ通信路に接続されて画像情報に基づく読話認識を実行する認識処理論理回路とを備える。したがって、音響情報を送信する送話器にカメラが取り付けられているため、送話器を使用する話者の顔がカメラに映し出され、読話のための顔の特徴に対応する画像情報が得られる。そこで、この画像情報がデータ通信路を介して認識処理論理回路に送信され、画像情報に基づく読話認識が実行される。

【0010】ここで、送話器は、例えば、ハンドセットやヘッドセットによって構成され、カメラは、例えば、デジタルカメラによって構成されている。カメラにより得られる画像情報は、例えば、使用者の舌の位置、唇の丸くふくらむ出っ張り、あごの位置である。あごの位置は、例えば、歯の隙間に基づく。

【0011】また、請求項1記載の発明は、電話装置に取り付けられて使用者の口元を照らす光源や赤外線光源を更に含んでいても良く、赤外線光源を含む場合、カメラは赤外線反応カメラや光学カメラによって構成される。

【0012】請求項2記載の発明は、画像情報に基づく読話認識を実行する認識処理論理回路で処理されるデータを得るシステムであり、送話器と、この送話器に組み合わされて読話のための少なくとも一つの顔の特徴に対応する画像情報を得るカメラと、このカメラに接続されてカメラから出力された画像情報を認識処理論理回路に送信するデータ通信路とを備える。したがって、音響情報を送信する送話器にカメラが取り付けられているため、送話器を使用する話者の顔がカメラに映し出され、読話のための顔の特徴に対応する画像情報が得られる。そこで、この画像情報がデータ通信路を介して認識処理論理回路に送信され、画像情報に基づく読話認識が実行される。

【0013】ここで、カメラにより得られる画像情報は、例えば、あごの位置であり、これは歯の隙間に基づく。カメラは、例えば、デジタルカメラによって構成されている。

【0014】また、請求項2記載の発明は、電話装置に取り付けられて使用者の口元を照らす光源や赤外線光源を更に含んでいても良く、赤外線光源を含む場合、カメラは赤外線反応カメラや赤外線光学反応カメラによって構成される。

【0015】

【発明の実施の形態】本発明の実施の形態を図面に基づいて説明する。

【0016】〔システムの概略〕図1は、電話機のハンドセット100の一例を示す。ハンドセット100は、カメラ101及び照明光源102を備える。カメラ101及び照明光源102は、ハンドセット100の送話器104の部分に対応するハウジング105（送話器ハウジングを兼ねる）に取り付けられている。照明光源10

2は、使用者（図示せず）が電話で話をしている間、使用者の口元を照明する。照明される領域は、カメラ101によって撮影される。カメラ101によって撮影された映像データは、データ通信路103経由で認識処理システムに送信され、認識を受ける。

【0017】電話機のハンドセット100は、標準形の電話機のハンドセットによって構成されている。もっとも、ハンドセットではなく、電話機のヘットセットとして形成されていても良い。

【0018】カメラ101は、ハンドセット100に直接的に取り付けられ、映像による音声認識に用いられるであろう情報を得るために、顔の特徴を抽出する映像データを獲得する。カメラ101は、小型のデジタルカメラによって構成されている。このようなカメラ101は、赤外線（IR）反応カメラ又は光学的カメラ（又は赤外線光学反応カメラ）である。

【0019】照明光源102は、赤外線（IR）光源によって構成されており、話者の口元を照明する。もっとも、存在している光（例えば、周辺光等）が話者の口元を照明するにの十分である場合には、照明光源102を作動させなくても良い。

【0020】データ通信路103は、処理及び分類のために画像情報を局所的な場所に送信する広い帯域幅（例えば、映像）のデータ通信路によって構成されている。このようなデータ通信路103は、また、ハンドセット100の送話器104に取り付けられたカメラ101によって撮影された画像データを、通信ネットワークやそれ自体が認識を受けるシステムに送信するよう構成されていても良い。

【0021】本実施の形態では、データ通信路103により送信される映像（及び音響）データは、少なくとも一つの認識アルゴリズムを受ける。認識は、ハンドセット100で受信された映像及び音響双方のデータについて実行され、これにより、より正確な認識結果を得る。

【0022】カメラ101の位置決めは非常に重要である。カメラ101は、0．5～5cm程度の幅を持っている。読話のために使用する必要な顔の特徴を得るために、カメラ101は、得られる映像データが真正面からの眺めとならないようにハンドセット100に位置決めされている。つまり、カメラ101は、話者の真正面の眺めに対してある角度をなす位置から話者の口元を撮影する。カメラ101の位置決めは、舌の位置（for/la/, /ta/, 他）及び唇の丸くふくらむ出っ張り（for/oo）を得ることができる真正面からの眺めに対してある角度をなす位置でなされる。ここでいう「ある角度」は、個々の使用者の顔の形状や個々の使用者がハンドセット100を耳に当てる角度等（特徴）に依存する。したがって、それらの特徴の検出、抽出が可能であるため、認識率の向上が期待できる。さらに、カメラ101の位置は、話者の歯の隙間を撮影することを許容するような位置でもある。これは、あごの位置を映像から直接検出することは非常に困難である反面、あごの位置は歯の隙間から確実に推察されるためである。したがって、本実施の形態のシステムは、舌の位置、唇の丸くふくらむ出っ張り、及び歯の隙間を使用して読話を実行する。もっとも、本発明は、それらの三つの特徴を使用するものには限定されず、他の顔の特徴を使用するようなものとして構成されていても良い。但し、使用可能な顔の特徴は、話者の口元に対するカメラ101の位置決め及び配置により限定される。

【0023】〔システムの詳細〕カメラ101からの入力データは、データ通信路103を経由して読話認識を実行する認識処理論理回路としての認識処理サブシステムに送信される。認識処理サブシステムは、数多くある周知のパターンマッチング技術を用いてパターンマッチングを実行する。例えば、認識処理サブシステムは、時間正規化（DTW：Dynamic Time Warping）パターン認識、隠れマルコフ・モデル（HMM：Hidden Markov Model）パターン認識、時間遅延ニューラル・ネットワーク（TDNN：Time Delay Neural Network）パターンマッチング、その他の認識処理技術を用いてパターンマッチングを実行する。

【0024】認識処理サブシステムは、また、読話認識と協力して音声認識を実行する。この方法では、読話認識は音声認識の正確さを高めるように動作する。

【0025】図2は、図1に示す入力装置を使用する模範的な読話認識システムのブロック図である。この認識システムは、システムバス201、中央処理装置（CPU）202、及びシステムメモリ203を中心として構成されている。認識される話者の口元は、照明光源102（図1参照）か、あるいは、オフィス環境で普通に得られるような通常の周辺光により照明される。映像は、例えば図1のカメラ101のような標準的なデジタルカメラであるビデオカメラ205によって記録され、出力されたラスタスキャン映像は、アナログデジタル変換器（ADC）204に送信される。このADC204では、システムメモリ203に格納する標準化及び量子化されたラスタイメージ（フレーム）を生成する。ラスタスキャンされた映像フレームのシーケンスは、ビデオカメラ205及びADC204によって処理され、話者による一又はそれ以上の発話を表現する。

【0026】ビデオカメラ205は、1秒間に30フレームを生成する。ADC204によって変換された後の各フレームは、640×480画素のアレイとなり、各画素は、ADC204により標準化された各点で映像の強度（輝度又はグレースケール）を表現する8ビットの数となる。各フレームの二つ組みのフィールドでは冗長度が高いために、フィールドは一つ置きに処分される。

【0027】システムメモリ203に格納された画素フレームは、空間周波数フィルタ206及び時間周波数フ

ィルタ207によって前処理される。空間周波数フィルタ206は、空間周波数ノイズを減少させるためにスムージング動作又は低域通過沪過動作を実行し、映像の輪郭をはっきりさせるためにエッジ強調動作を実行する。空間周波数沪過された映像は、また、三つのシーケンシャルフレームの幅で円滑化、すなわち、時間周波数フィルタ207の低域通過フィルタを用いる時間周波数スムージングがなされる。沪過動作が第一の又は組み合わせのスムージングとして実行されるか、エッジシャープニングが単一の沪過動作として実行される。空間周波数沪過及び時間周波数沪過は、周知の技術である。システムメモリ203には沪過された映像が格納される。

【0028】沪過された映像が得られたなら、後続する処理に使用される映像のサイズを縮小すること、すなわち、発音された発話情報を含む関心領域（ROI）だけを保つことが望まれる。ROIは、口の開きに集中する。

【0029】口元以外の顔の部分は、口の動きに比べてフレーム間で静止（固定）している傾向が強く、連続的なフレーム間の変化は口元（ROI）に多い。明るい画素は、後続するフレーム間の大きな変化の点、すなわち、口を連想する画素でありそうなものを表現している。

【0030】ROIは、映像データに適用されるマスク作用によって限定される。マスクは、グレースケールの閾値によって作成される。マスクを使用し、空間周波数領域は、顔の特徴が得られた場所から特定される。この領域は、三つの時空間座標によって定義される長方形領域又は重心領域よりなる。各フレームのために、二つの空間座標だけが使用されることに留意されたい。格納された映像に対応する領域が取り入れられ、これによって、ROIによって限定された範囲に含まれている映像画素だけが格納される。

【0031】そして、与えられた発話に関連付けられるROIの収集した一揃えは、時間指数の関数nとしてy軸のある地点でグレースケールから切り取られる。それは、顔の特徴の抽出のために使用されるかもしれない一揃えのきっかけを形成するために用いられる情報である。以前に確定された顔の特徴の抽出は、特徴抽出ユニット209によって実行される。

【0032】顔の特徴が抽出されたなら、ソフトウエアに従い動作するCPU202によってパターン認識が実行される。DTWの場合、参照辞書に数多くの周知の参照パターンが格納されている。DTW処理の間、未知（入力映像）パターンと参照パターンとの間のマッチングをとるために、未知パターンが辞書の参照パターンと比較される。未知パターンの各ポイントは、各参照パターンの各ポイントと比較され、左から右にラティス（格子形データ）を横切って走る点のラティス又はアレイの幅でコスト関数が生成される。コストというのは、未知パターンと参照パターンとの間の距離である。DTW処理の目標は、各参照パターンのための格子形データを横切って最も低いコストのパスを探し当て、未知パターンに最もマッチするパターンを探し当てるために、参照パターンのそれぞれのパスを比較することである。最もマッチするパターンは、読話認識の結果としてシステムから出力される。

【0033】以上述べた通り、本実施の形態では、電話機のハンドセット（又はヘッドセット）に直接据え付けられているカメラを使用し、読話認識に用いるための顔の特徴を得る。そして、音響情報に基づく音声認識を補強するように画像情報に基づく読話認識を用いることで、認識精度を維持したまま、より多くの語彙についてより多くの人に適用できる音声認識を行うことができるようになる。

【0034】ここで、読話認識は、例えば、放送システムで使用するための音声音響認識とともに実行されるかもしれないことに留意されたい。このような場合、音声データは、また、音声データが映像データとは全く違ったものとして受信され沪過されることを除き、上記と類似の方法で認識される。映像及び音声データの双方は、後の認識結果の収集のために時間が切り取られても良い。このような場合、認識結果は、得られた映像及び音声データと最もマッチする顔の特徴と音声の特徴との双方を備える参照パターンである。

【0035】本発明は、（対話よりもむしろ）単一の話者の話を録音・再生することを目標とするようなものに有効に利用されるであろう。そのような適用対象の例としては、株取引のような金融取引、例えば会社の購買部等に設置される電話注文の自動録音等が含まれる。

【0036】本発明の優位点の一つは、確かな取引を提供できるということである。なぜなら、単一の電話機は概して同一の話者に繰り返し使用され、読話認識システムがトレーニングされれば信頼性が高まるからである。更に、本発明の別の優位点は、認識アルゴリズムは話者の話の長期間に渡る変化を良く追跡することができるに違いないということである。

【0037】本発明の別の優位点は、顔の特徴の撮影及びその後に続く認識を実行するために必要な構成要素として、読話システム、デジタルカメラ、並びに広帯域な通信及びプロトコルというような通常のものを用いることができる、ということである。

【0038】上述の記述を読んだ後の当業者にとって、本発明の多くの変更や修正が明かになることは疑いないであろうし、図面として示し説明した特定の態様は、発明内容の限定を意図したものではない。実施の態様は特許請求の範囲を限定することを意図したものではない。

【0039】

【発明の効果】本発明は、読話のための少なくとも一つの顔の特徴に対応する画像情報を得るカメラを送話器に

据付け、カメラから出力される画像情報に基づいて読話認識を実行できるようにしたので、音響情報に基づく音声認識を補強するように画像情報に基づく読話認識を用いることで、認識精度を維持したまま、より多くの語彙についてより多くの人に適用できる音声認識を行うことができる。

【図面の簡単な説明】

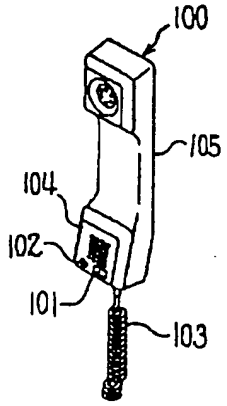
【図1】本発明の実施の一形態として、電話機のハンドセット（送話器）の外観斜視図を示す。

【図2】本発明の実施の一形態として、画像認識システムのブロック図を示す。

【符号の説明】

105	送話器ハウジング
104	送話器
101，205	カメラ
103	データ通信路

【図1】



【図2】